DLRG@DravidianLangTech-ACL2022: Abusive Comment Detection in Tamil using Multilingual Transformer Models

Ankita Duraphe and **Ratnavel Rajalakshmi***
School of Computer Science and Engineering
Vellore Institute of Technology
Chennai, India
ankitaduraphe@gmail.com
rajalakshmi.r@vit.ac.in

Antonette Shibani
TD School
University of Technology Sydney
Sydney, Australia
antonette.shibani@gmail.com

Abstract

Online Social Network has let people connect and interact with each other. It does, however, also provide a platform for online abusers to propagate abusive content. The majority of these abusive remarks are written in a multilingual style, which allows them to easily slip past internet inspection. This paper presents a system developed for the Shared Task on Abusive Comment Detection (Misogyny, Misandry, Homophobia, Transphobic, Xenophobia, CounterSpeech, Hope Speech) in Tamil DravidianLangTech@ACL 2022 to detect the abusive category of each comment. We approach the task with three methodologies - Machine Learning, Deep Learning and Transformer-based modeling, for two sets of data - Tamil and Tamil+English language dataset. The dataset used in our system can be accessed from the competition on CodaLab. For Machine Learning, eight algorithms were implemented, among which Random Forest gave the best result with Tamil+English dataset, with a weighted average F1-score of 0.78. For Deep Learning, Bi-Directional LSTM gave best result with pre-trained word embeddings. In Transformer-based modeling, we used IndicBERT and mBERT with fine-tuning, among which mBERT gave the best result for Tamil dataset with a weighted average F1-score of 0.7.

1 Introduction

The usage of the Internet and social media has increased exponentially over the previous two decades, allowing people to connect and interact with each other (Priyadharshini et al., 2021; Kumaresan et al., 2021). This has resulted in a number of favourable outcomes such as monitoring pandemic trends, empowering patients and enhancing public communication through social media, amongst others (Cornelius et al., 2020; Househ et al., 2014; Picazo-Vela et al., 2012). At the same time, it has also brought with it hazards and negative consequences, one of which is the use of abusive language on others (Chakravarthi, 2020; Chakravarthi and Muralidaran, 2021).

The rapid spread of abusive content on social networking has become a major source of concern for government organisations. It is very difficult to identify abuse over online social network due to the massive volume of content generated through social media in different online platforms (Sampath et al., 2022; Ravikiran et al., 2022; Chakravarthi et al., 2022; Bharathi et al., 2022). It becomes a bigger problem when most of the communication is in multilingual style (Priyadharshini et al., 2020; Chakravarthi et al., 2021a,b). Hence, there is increasing interest in the use of automated methods for detecting online social abuse (Priyadharshini et al., 2022). It is becoming a major area of research to find solutions with powerful algorithmic systems to curb the growth of abusive content online. One possible way of achieving such a system is by using state-of-the-art Natural Language Processing (NLP) techniques, which can analyse, comprehend and interpret the meaning of the natural language data.

In addition, the detection of abusive language online is harder for some languages like Tamil due to the presence of code-mixed (Barman et al., 2014) and code-switched (Poplack, 2001) data. Code-switching is when in a single discourse, a person switches between two or more languages or language varieties/dialects (B and A, 2021b,a). It refers to using elements from more than one language in a way that is consistent with the syntax, morphology, and phonology of each language or dialect. Code-mixing is the hybridization of two languages (for example, parkear, which uses an English root word and Spanish morphology), which refers to the migration from one language to another. Many such language pairs have a hybrid

*Corresponding Author

Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, pages 207 - 213

name.

Tamil is a member of the southern branch of the Dravidian languages, a group of about 26 languages indigenous to the Indian subcontinent. It is also classed as a member of the Tamil language family, which contains the languages of around 35 ethno-linguistic groups, including the Irula and Yerukula languages (Anita and Subalalitha, 2019b,a; Subalalitha and Poovammal, 2018; Subalalitha, 2019). Malayalam is Tamil's closest significant cousin; the two began splitting during the 9th century AD. Although several variations between Tamil and Malayalam indicate a pre-historic break of the western dialect, the process of separating into a different language, Malayalam, did not occur until the 13th or 14th century (Sakuntharaj and Mahesan, 2021, 2017, 2016; Thavareesan and Mahesan, 2019, 2020a,b, 2021). Tanglish is an example which is Tamil+English. In this task, we are given two datasets: One with a Tamil meaning written in English but the content is a combination of Tamil and English. The other is a Tamil+English dataset (Tanglish) which is written in Tamil and English with content in Tamil and English as well. There are also known challenges in the development of computational systems in Tamil because of the lack of linguistic resources (Magueresse et al., 2020). In this paper, we present computational systems for the automated detection of abusive language using the two different data sets containing Tamil and Tamil+English.

2 Related work

In this section, we review the various methodologies and systems previously implemented for similar tasks in under-resourced languages like Tamil. Hope speech is annotated Equality, Diversity and Inclusion (HopeEDI) (Chakravarthi, 2020). They also created several baselines to standard the dataset. (Chakravarthi and Muralidaran, 2021) reports on the shared task of hope speech detection for Tamil, English and Malyalam languages. They presents the dataset used in the shared task and also surveys various competing approaches developed for the shared task and their corresponding results. (Mandalam and Sharma, 2021) presents the methodologies implemented while classifying Dravidian Tamil and Malayalam code-mixed comments according to their polarity and uses LSTM architecture. (Sai and Sharma, 2021; Li, 2021; Que, 2021) use XLM-RoBERTa for offensive language identification. Novel approach of selective translation and transliteration have been used to improve the performance of multilingual transformer networks such as XLMRoBERTa and mBERT by fine-tuning and ensembling. Online messaging has become one of the most popular methods of communication with instances of online/digital bullying. The challenge of detecting objectionable language in YouTube comments from the Dravidian languages of Tamil, Malayalam, and Kannada is viewed as a multi-class classification problem (Andrew, 2021). Several Machine Learning algorithms have been trained for the task at hand after being exposed to language-specific pre-processing.

3 Dataset

The dataset for the current study is taken from the competition [1] which consists of YouTube comments in Tamil and Tamil-English languages annotated for Misogyny, homophobia, transphobic, xenophobia, counter-speech, hope-speech and misandry (and None-of-the-above) (Priyadharshini et al., 2022). Table 1 shows the count of comments for both the datasets under each split. Table 2 gives the class-distribution of each abusive category for both the datasets.

4 Proposed Technique

Raw texts are inaccessible to Machine Learning (ML) and Deep Learning (DL) algorithms. To train the models for classification, feature extraction is necessary. To extract features in ML approaches, the TF-IDF representation is used. For DL models, we use fastText word embeddings feature extraction strategies (Joulin et al., 2016). fastText embedding uses a pre-trained embedding matrix for Tamil language (Grave et al., 2018). To study the results and come up with the best model possible, we follow three approaches - Machine Learning, Deep Learning and Transformer-based.

As it can be clearly seen from Table 1, both the datasets contain class imbalance. Class imbalance is a problem in machine learning when there are great differences in the class-distribution of the dataset. It is seen as a problem when a dataset is biased towards a class in the dataset. If this problem persists, any algorithm trained on the same data will again be biased towards the same class. To resolve

[1] https://competitions.codalab.org/competitions/36403

Class	Tamil+English	Tamil
Train-set	5948	2238
Validation-set	1488	560
Test-set	1857	699

Table 1: Number of comments across both the datasets in each of the three splits.

Class	Tamil+English	Tamil
Misandry	1048	550
Counter-speech	443	185
Xenophobia	367	124
Hope-Speech	266	97
Misogyny	261	149
Homophobia	213	43
Transphobic	197	8
None-of-the-above	4639	1642

Table 2: Class-distribution across both datasets.

the issue of class imbalance, we practice various approaches:

Changing the performance metric: Since accuracy is not always the best metric to use on imbalanced datasets, we use F1-score instead to evaluate the models.

Using a penalized algorithm (cost-sensitive training): This algorithm also handles class imbalance which can be achieved by using 'balanced' as a parameter while computing class weights.

Changing the algorithms: This is why we have used a wide variety of algorithms to get a bigger picture of which models suit the dataset and the classification problem better.

Table 3 provides the details about tuning the hyperparameters in our system both for Tamil+English and Tamil datasets.

To study the results and come up with the best model possible, we follow three approaches - Machine Learning, Deep Learning and Transformer-based, described in the sub-sections below.

4.1 Approach A: Machine Learning/ Non-Neural Network approaches

To start with, we implemented various Machine Learning algorithms which include Logistic Regression (LR), Random Forest (RF), K-nearest neighbors (KNN), Decision Tree, Support Vector Machine (SVM), Gradient Boosting, Adaptive Boosting (AdaBoost), and Ensemble (Husain, 2020). We have used ML algorithms only for Tamil+English dataset due to the poor performance of ML models on Tamil written text (Tamil dataset).

4.2 Approach B: Recurrent Neural Network approaches

To improve the performance of ML models, we dive into deep learning algorithms. Here, we have implemented DL approach for both the datasets. We use two models of Bi-directional LSTM - BiLSTM-M1 and BiLSTM-M2 (Chiu and Nichols, 2015). BiLSTM-M1 is a mix of bidirectional LSTM architecture that uses a convolution and a max-pooling layer to extract a new feature vector from the per-character feature vectors for each word. These vectors are concatenated for each word and sent to the BiLSTM network, which subsequently feeds the output layers. BiLSTM-M2 is an advanced BiLSTM-M1 where we adopted pre-trained word embeddings since BiLSTM and fastText produced better results for classification tasks.

4.3 Approach C: Transformer-based approaches

In natural language processing, the Transformer is a unique design that seeks to solve sequence-to-sequence tasks while also resolving long-range dependencies. It does not use sequence-aligned RNNs or convolution to compute representations of its input and output, instead relying solely on self-attention.

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) is a

Parameters	Values
Learning rate	$1x10^{-3}$
Batch Size	32
Epochs	25
Validation Split	0.2

Table 3: Hyperparameters used in our system.

Model name	P	R	F1
RF	**0.91**	**0.71**	**0.78**
Gradient Boosting	0.85	0.71	0.76
SVM	0.78	0.72	0.75
KNN	0.85	0.68	0.75
AdaBoost	0.86	0.69	0.74
LR	0.71	0.71	0.71
Decision Tree	0.72	0.66	0.68
Ensemble	0.71	0.72	0.68
BiLSTM-M1	0.71	0.68	0.7
BiLSTM-M2	0.64	0.61	0.62
IndicBERT	0.55	0.67	0.60

Table 4: Metric evaluation for Tamil+English dataset.

Model name	P	R	F1
BiLSTM-M1	0.63	0.55	0.58
BiLSTM-M2	**0.74**	**0.67**	**0.7**
mBERT	0.64	0.7	0.7

Table 5: Metric evaluation for Tamil dataset.

transformer language model with a variable number of encoder layers and self-attention capabilities.

We again use two BERT models - mBERT (bert-base-multilingual-cased) and IndicBERT

We follow fine-tuning for Transformer models and use pre-trained BERT, bert-base-multilingual-cased (Devlin et al., 2018) and IndicBert classification models (Kakwani et al., 2020) that have been trained on 104 languages and 12 Indian languages respectively, including Tamil, from the largest Wikipedia.

5 Results and Discussion

We ran 8 Machine Learning algorithms, 2 Deep Learning and 1 Transformer model on the Tamil+English dataset. For the Tamil dataset, we used 2 Deep Learning and 1 Transformer model.

For the Tamil+English dataset, the best performance was of Random Forest with macro average F1-score of 0.32 and weighted average F1-score of 0.78. For the Tamil dataset, the best model was BiLSTM-M2 with macro average F1-score of 0.39 and weighted average F1-score of 0.70.

For Tamil, performance improved from switching BiLSTM-M2 to mBERT. And for Tamil+English, the best performer was BiLSTM-M1, followed by BiLSTM-M2 and then IndicBERT and mBERT.

Table 4 and Table 5 show the result of our models across both the datasets. For Tamil language, ML models performed best when DL models were originally expected to perform better. The extensive use of multilingual language in the text could be a reason for the poor performance of DL. Pre-trained word embeddings could not deliver higher performance due to the lack of feature mapping between the words. As a result, DL models might not be able to uncover sufficient relational relationships among the features, and perform poorly.

6 Conclusions and Future Work

In this paper, we presented approaches for the automated detection of abusive comments in Tamil. We used various models to do a comparative study to see which model performed better with the dataset given in the shared task. We found that Deep Learning and Transformer models outperformed Machine Learning models with Tamil data whereas Machine Learning models achieved better results than Deep Learning and Transformer-based for Tamil+English data. We did not apply contextualized embeddings (such as ELMO, FLAIR) which may improve the performance of the system. Implementation of Contextualised embeddings using language modelling with deep learning is the future work to explore.

Acknowledgements

We would like to thank the management of Vellore Institute of Technology, Chennai for their support to carry out this research.

References

Judith Jeyafreeda Andrew. 2021. JudithJeyafreedaAndrew@DravidianLangTech-EACL2021:offensive language detection for Dravidian code-mixed YouTube comments. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 169–174, Kyiv. Association for Computational Linguistics.

R Anita and CN Subalalitha. 2019a. An approach to cluster Tamil literatures using discourse connectives. In *2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP)*, pages 1–4. IEEE.

R Anita and CN Subalalitha. 2019b. Building discourse parser for Thirukkural. In *Proceedings of the 16th International Conference on Natural Language Processing*, pages 18–25.

Bharathi B and Agnusimmaculate Silvia A. 2021a. SSNCSE_NLP@DravidianLangTech-EACL2021: Meme classification for Tamil using machine learning approach. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 336–339, Kyiv. Association for Computational Linguistics.

Bharathi B and Agnusimmaculate Silvia A. 2021b. SSNCSE_NLP@DravidianLangTech-EACL2021: Offensive language identification on multilingual code mixing text. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 313–318, Kyiv. Association for Computational Linguistics.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media.

B Bharathi, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, N Sripriya, Arunaggiri Pandian, and Swetha Valli. 2022. Findings of the shared task on Speech Recognition for Vulnerable Individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi. 2020. HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

Bharathi Raja Chakravarthi and Vigneshwaran Muralidaran. 2021. Findings of the shared task on hope speech detection for equality, diversity, and inclusion. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 61–72, Kyiv. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John Phillip McCrae, Paul Buitaleer, Prasanna Kumar Kumaresan, and Rahul Ponnusamy. 2022. Findings of the shared task on Homophobia Transphobia Detection in Social Media Comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kayalvizhi Sampath, Durairaj Thenmozhi, Sathiyaraj Thangasamy, Rajendran Nallathambi, and John Phillip McCrae. 2021a. Dataset for identification of homophobia and transophobia in multilingual YouTube comments. *arXiv preprint arXiv:2109.00227*.

Bharathi Raja Chakravarthi, Priya Rani, Mihael Arcan, and John P McCrae. 2021b. A survey of orthographic information in machine translation. *SN Computer Science*, 2(4):1–19.

Jason P. C. Chiu and Eric Nichols. 2015. Named entity recognition with bidirectional lstm-cnns. *CoRR*, abs/1511.08308.

Joseph Cornelius, Tilia Ellendorff, Lenz Furrer, and Fabio Rinaldi. 2020. COVID-19 Twitter monitor: Aggregating and visualizing COVID-19 related trends in social media. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 1–10, Barcelona, Spain (Online). Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Avishek Garain, Atanu Mandal, and Sudip Kumar Naskar. 2021. JUNLP@DravidianLangTech-EACL2021: Offensive language identification in Dravidian langauges. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 319–322, Kyiv. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Adeep Hande, Ruba Priyadharshini, Anbukkarasi Sampath, Kingston Pal Thamburaj, Prabakaran Chandran, and Bharathi Raja Chakravarthi. 2021. Hope speech detection in under-resourced kannada language.

Mowafa Househ, Elizabeth Borycki, and Andre Kushniruk. 2014. Empowering patients through social media: the benefits and challenges. *Health Informatics J.*, 20(1):50–58.

Fatemah Husain. 2020. Arabic offensive language detection using machine learning and ensemble machine learning approaches. *CoRR*, abs/2005.08946.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Prasanna Kumar Kumaresan, Ratnasingam Sakuntharaj, Sajeetha Thavareesan, Subalalitha Navaneethakrishnan, Anand Kumar Madasamy, Bharathi Raja Chakravarthi, and John P McCrae. 2021. Findings of shared task on offensive language identification in Tamil and Malayalam. In *Forum for Information Retrieval Evaluation*, pages 16–18.

Zichao Li. 2021. Codewithzichao@DravidianLangTech-EACL2021: Exploring multilingual transformers for offensive language identification on code mixing text. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 164–168, Kyiv. Association for Computational Linguistics.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *CoRR*, abs/2006.07264.

Asrita Venkata Mandalam and Yashvardhan Sharma. 2021. Sentiment analysis of Dravidian code mixed data. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 46–54, Kyiv. Association for Computational Linguistics.

Sergio Picazo-Vela, Isis Gutiérrez-Martínez, and Luis Felipe Luna-Reyes. 2012. Understanding risks, benefits, and strategic alternatives of social media applications in the public sector. *Gov. Inf. Q.*, 29(4):504–511.

Shana Poplack. 2001. *Code Switching: Linguistic*, pages 2062–2065.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumar Kumaresan. 2022. Findings of the shared task on Abusive Comment Detection in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Sajeetha Thavareesan, Dhivya Chinnappa, Durairaj Thenmozhi, and Rahul Ponnusamy. 2021. Overview of the DravidianCodeMix 2021 shared task on sentiment detection in Tamil, Malayalam, and Kannada. In *Forum for Information Retrieval Evaluation*, pages 4–6.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Mani Vegupatti, and John P McCrae. 2020. Named entity recognition for code-mixed Indian corpus using meta embedding. In *2020 6th international conference on advanced computing and communication systems (ICACCS)*, pages 68–72. IEEE.

Qinyu Que. 2021. Simon @ DravidianLangTech-EACL2021: Detecting offensive content in Kannada language. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 160–163, Kyiv. Association for Computational Linguistics.

Manikandan Ravikiran, Bharathi Raja Chakravarthi, Anand Kumar Madasamy, Sangeetha Sivanesan, Ratnavel Rajalakshmi, Sajeetha Thavareesan, Rahul Ponnusamy, and Shankar Mahadevan. 2022. Findings of the shared task on Offensive Span Identification in code-mixed Tamil-English comments. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Siva Sai and Yashvardhan Sharma. 2021. Towards offensive language identification for Dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 18–27, Kyiv. Association for Computational Linguistics.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2016. A novel hybrid approach to detect and correct spelling in Tamil text. In *2016 IEEE International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 1–6.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2017. Use of a novel hash-table for speeding-up suggestions for misspelt tamil words. In *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*, pages 1–5.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2021. Missing word detection and correction based on context of Tamil sentences using n-grams. In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 42–47.

Anbukkarasi Sampath, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Ruba Priyadharshini, Subalalitha Chinnaudayar Navaneethakrishnan, Kogilavani Shanmugavadivel, Sajeetha Thavareesan, Sathiyaraj Thangasamy, Parameswari Krishnamurthy, Adeep Hande, Sean Benhur, and Santhiya Ponnusamy, Kishor Kumar Pandiyan. 2022. Findings of

the shared task on Emotion Analysis in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

C. N. Subalalitha. 2019. Information extraction framework for Kurunthogai. *Sādhanā*, 44(7):156.

CN Subalalitha and E Poovammal. 2018. Automatic bilingual dictionary construction for Tirukural. *Applied Artificial Intelligence*, 32(6):558–567.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2019. Sentiment analysis in Tamil texts: A study on machine learning techniques and feature representation. In *2019 14th Conference on Industrial and Information Systems (ICIIS)*, pages 320–325.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020a. Sentiment lexicon expansion using Word2vec and fastText for sentiment prediction in Tamil texts. In *2020 Moratuwa Engineering Research Conference (MERCon)*, pages 272–276.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020b. Word embedding-based part of speech tagging in Tamil texts. In *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, pages 478–482.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2021. Sentiment analysis in Tamil texts using k-means and k-nearest neighbour. In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 48–53.